This paper was accepted for presentation at the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28, 2005.

Predicting the importance of current papers

Kevin W. Boyack (correspondence author), Sandia National Laboratories, P.O. Box 5800, MS-0310, Albuquerque, NM 87185, USA email:kboyack@sandia.gov

Richard Klavans, SciTech Strategies, Inc., Berwyn, PA, 19312, USA
email: rklavans@mapofscience.com

Abstract: This article examines how well one can predict the importance of a current paper (a paper that is recently published in the literature). We look at three factors – journal importance, reference importance and author reputation. Citation-based measures of importance are used for all variables. We find that journal importance is the best predictor (explaining 22.3% out of a potential 29.1% of the variance in the data), and that this correlation value varies significantly by discipline. Journal importance is a better predictor of citation in Computer Science than in any other discipline. While the finding supports the present policy of using journal impact statistics as a surrogate for the importance of current papers, it calls into question the present policy of equally weighting current documents in text-based analyses. We suggest that future researchers take into account the expected importance of a document when attempting to describe the cognitive structure of a field.

Background

Bibliometrics has spent much of the last twenty years as an outsider in terms of its role in the research evaluation process. This has changed in the last few years as bibliometrics has become a fashionable partner to the process in many circles. It is becoming a more common policy (despite the related controversy) to use journal quality in the form of the impact factor for evaluating the importance of the recent work of a faculty member, a department, or even an entire organization (Glänzel & Moed, 2002). The misuse of citation-based indicators in research evaluation processes has recently sparked commentary from those who are proficient in their generation and use. For instance, Weingart (2005) and van Raan (2005) comment expertly on potential pitfalls, and the care and techniques needed to avoid them.

It is generally accepted that citation counts are a reasonable indicator of the importance of a scientific paper. However, citations at the paper level are rarely used in research evaluation due to the time factor; evaluation typically focuses on the most recently published work, and these very young papers often haven't had enough time to accrue sufficient citations for analysis. Few studies have been done to predict future citation rates. However, these show that reasonable predictions of future citation counts are possible from the citation counts of a shorter time period (e.g. 3-5 years) following publication (Glänzel, 1997; Glänzel & Schubert, 1995).

The fall-back position, given time factors, has been to use journal impact factors as proxy for actual paper citations counts under the assumption that, in the aggregate, future citations are a function of journal quality. Numerous authors counsel against this (cf., van Raan, 2005). However, due to the ease of using impact factors, it is likely that they will continue to be used (and misused) despite arguments to the contrary. Thus, we feel a study showing the accuracy of impact factors as predictors of paper-level citation statistics is timely. In this paper we explore the impact of several factors, including journal impact factors, on short-term citation rates. We also discuss how such factors can be used to create more accurate maps of science for all of science or for specific disciplines.

Dependent and Independent Variables

In order to study the effects of different variables on the number of times a paper is cited, we have constructed a set of data from combined SCIE/SSCI for the years 2002 and 2003. Of the

1.07 million individual records available in the 2002 fileyear, we limited our analysis to 780,049 papers that a) were bibliographically coupled to at least one other paper in the set, and b) were from a journal with a 2002 impact factor. Thus, our filter, while it does allow editorials with significant reference lists, is also more limiting than the commonly used ALNR (articles, letters, notes, reviews) filter. Any ALNR that do not meet the bibliographic coupling criteria are excluded from analysis. These 780,049 papers were defined as *current papers.*

Here, we investigate the influence of three independent variables on our dependent variable, CITED, defined as the number of times *current papers* were mentioned in the reference lists of articles indexed in 2002 and 2003. Papers indexed early in 2002 had nearly 2 years to accrue citations (from the beginning of 2002 to the end of 2003), while those indexed near the end of 2002 had only 1 year to accrue citations. On average, each paper had 1.5 years to accrue citations. This is an admittedly small citation window following publication. However, many of the papers considered as part of a research evaluation exercise are less than three years old, justifying use of a short citation window for this study. The total number of citations accrued by the *current papers* by the end of 2003 was 1,534,432. 52% of the *current papers* received at least one citation by the end of 2003, while 48% of the papers remained uncited.

The three independent variables considered are 1) journal importance (hereafter JIMPACT), 2) reference impact (hereafter REFIMPACT), and 3) author reputation (hereafter AUTHREP). The first independent variable was the importance of the journal. We used the 2002 journal impact factor calculated from the raw 2002 citation data using the journal impact factor formula published by ISI (2002 citations to journals in 2000/2001 divided by the number of papers in the 2000/2001 issues of those journals). The journal impact factor can be thought of simplistically as a two-year average citation rate. There were 7335 journals associated with the *current papers* in 2002. Journal impact ranged from 0.003 to 50.5, with an average value of 2.04. Note that our journal impact numbers will vary slightly from those published in the JCR, primarily because of algorithmic differences in matching of references with previously indexed papers. Our matching algorithm is undoubtedly different than the one used by ISI. We also realize there is a slight temporal incongruence between using JIMPACT based on citations from 2002 to the previous two years, and using CITED based on citations from 2002 and 2003 to *current papers*, but it cannot be avoided given the nature of this study.

Figure 1 shows the distributional characteristics of CITED (the number of times an article was cited in the future) and JIMPACT (the journal impact factor). The line in Figure 1 represents the tendency for CITED to increase as JIMPACT increases. 95% of all observations are below this line. For example, it is extremely unlikely that a paper that is published in a journal with an impact factor of 1 will be cited 10 times (this point is above the line, and fewer than 1% of the papers in this journal impact range are cited more than 10 times). But it is very likely that a paper that is published in a journal with an impact factor of 10 will be cited more than 10 times (this point is below the line, and over 42% of the papers in this journal impact range are cited more than 10 times). One can expect that CITED and JIMPACT will be highly related.

The second independent variable was reference impact, where reference impact was calculated as the number of times the references (those indexed by ISI) of a particular *current paper* were cited by all 2002 *current papers*. We explored the possibility that an article with references that were highly cited was an article that would be cited more highly in the future.

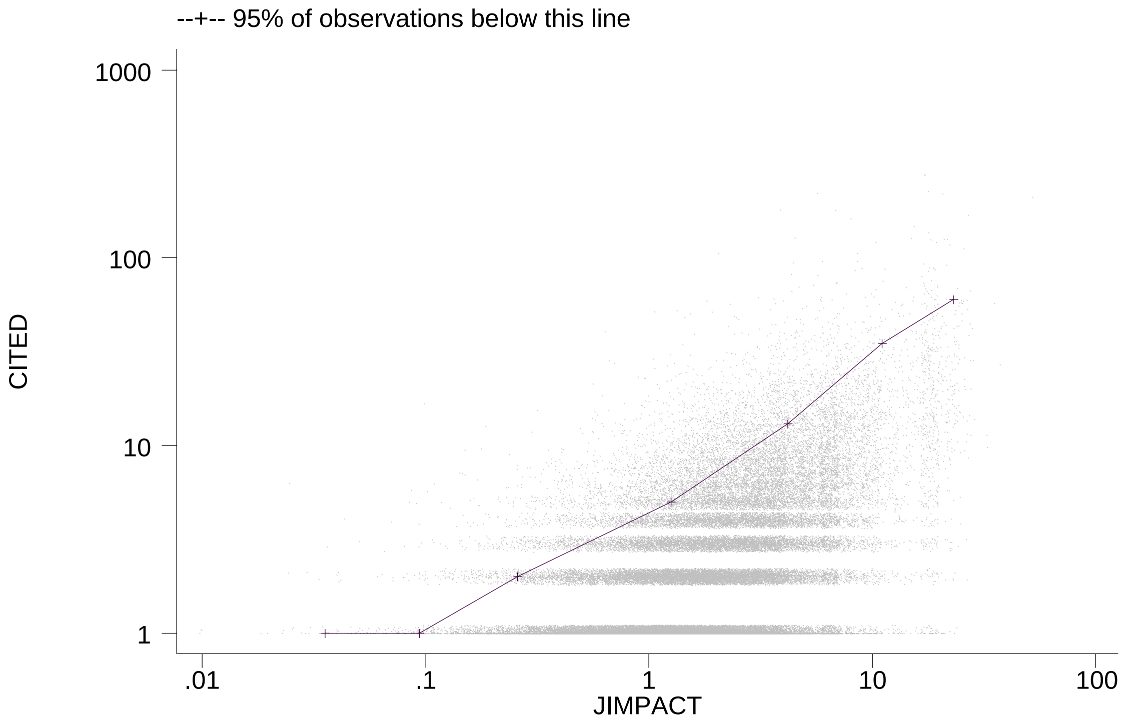

Figure 1: Relationship between CITED and JIMPACT (100,000 points sampled, points are dithered to illustrate density).

This gives more weight to review articles (which tend to be cited more) and articles with a well thought-out bibliography (the authors select the more highly cited references). Articles with no bibliography (or with a bibliography that wasn't cited by other articles in 2002) were assigned a reference impact of zero. The average value was 934. 97.5% of the papers in this sample had a positive reference impact value; only 2.5% of the papers had references that weren't cited.

Figure 2 shows the relationship between CITED and REFIMPACT (the number of citations to the references in the article). The line in Figure 2 suggests that CITED increases as REFIMPACT increases. 95% of all observations are below this line. This relationship is similar to the relationship in Figure 1. At first glance, this is not surprising, since articles naturally tend to cite at least some articles in the same journal. One can reasonably expect there to a relationship between JIMPACT and REFIMPACT. However, the self-citation rate is not particularly high, and does not contribute to a perceived correlation between JIMPACT and REFIMPACT as much as might be assumed. Figure 3 shows the relationship between self-citation fraction and JIMPACT. The self-citation fraction over all journals is 0.135, and decreases with increasing journal impact. Thus, self-citation does not explain correlation between JIMPACT and REFIMPACT at the high end. It is more likely that articles in a high impact journal tend to reference articles in other related high impact journals. By contrast, articles in low impact journals tend to reference articles in the same (low impact) journal and in other related lower impact journals.

The third independent variable was author reputation. We used four major assumptions to constrain our calculation of a measure of author reputation. First, we assumed that a single author might have a different reputation in different journals, especially if the author

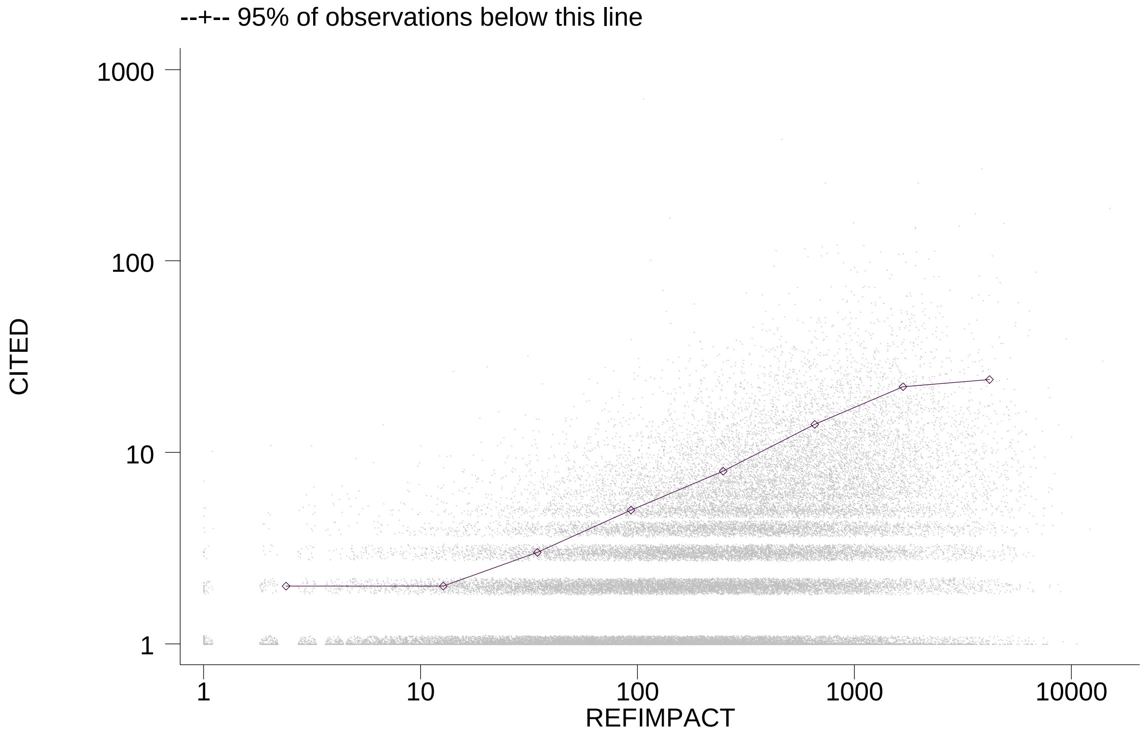

Figure 2: Relationship between CITED and REFIMPACT (100,000 points sampled, points are dithered to illustrate density).

publishes in more than one discipline. To capture this phenomenon we calculated reputation by author by journal. We did not specifically try to address the issue of author name [dis]ambiguation, either due to spelling variances or homography. However, limiting our calculation to author-journal pairs reduces the problem of ambiguous author names because fewer authors with the same name publish in the same journal than across all of science. Second, we used all authors, not just first authors, and counted a full paper-journal count for each author. Third, we limited the matching of author-journal pairs to papers published in the previous four years, namely 1998-2001. This was done to get a measure of recent author reputation. And fourth, we removed all editorials, news articles, book reviews, and all other non-technical records from consideration. This last step removed nearly all of the very large reputation numbers, which otherwise would have gone to journal editors and news writers. 66.3% of the papers had authors that had appeared in the literature in the past 2 years. One author was an author for 100 articles in the same journal during this time period.

Figure 4 shows the relationship between CITED and AUTHREP (where CITED>0 and AUTHREP>0), where AUTHREP is the number of times an author-journal pair from the 2002 *current papers* occurred in 1998-2001. From Figure 4, it does not appear that author reputation has the same strong positive association with future citation rate as does journal impact factor or reference impact. The line representing 95% of the observations (by AUTHREP) has a very minor (but positive) slope, suggesting a relatively weak relationship between AUTHREP and CITED. It is very possible that using citation numbers rather than simple paper counts to measure author reputation might have a stronger positive correlation. We plan to investigate this in the future.

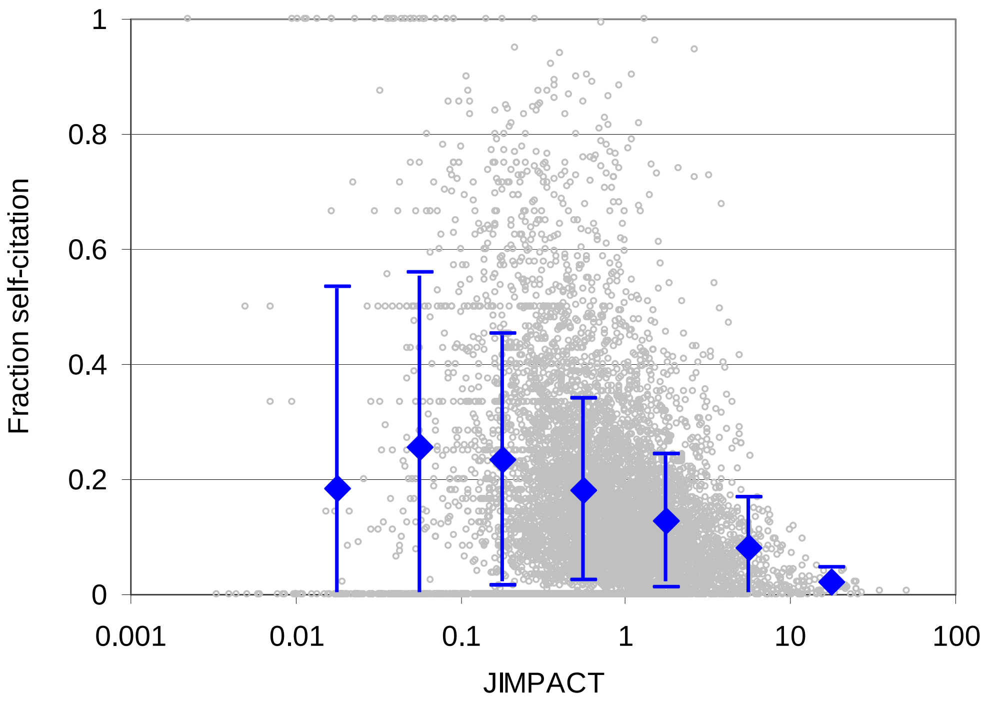

Figure 3: Relationship between JIMPACT and fraction self-citation for 7300 journals. Diamonds represent the average self-citation fractions by range. Bars represent one standard deviation.

Correlation and Regression Analysis

Values of CITED, JIMPACT, REFIMPACT, and AUTHREP were calculated for each of the 780,049 current papers. Log transforms were applied to the data to help deal with skewness. The following equations were used to incorporate the large number of observations with values of zero.

LCITE = log (CITED + 1)
LJOUR = log (JIMPACT) (there were no journals with a zero impact factor)
LREF = log (REFIMPACT + 1)
LAUT = log (AUTHREP + 1)

The correlation between these four variables (Table 1) suggests that the journal impact factor is most important (the highest Pearson correlation in column 1). However, the relationship between journal impact and reference impact is exceptionally high (.5091). This high correlation is consistent with the comment made above (papers in high impact journals will have high impact references).

Table 1: Correlation matrix between variables in this study.

	LCITE	**LJOUR**	**LREF**	**LAUTH**
LCITE	1.0000	0.4777	0.4483	0.1719
LJOUR	0.4777	1.0000	0.5091	0.1518
LREF	0.4483	0.5091	1.0000	0.1574
LAUTH	0.1719	0.1518	0.1574	1.0000

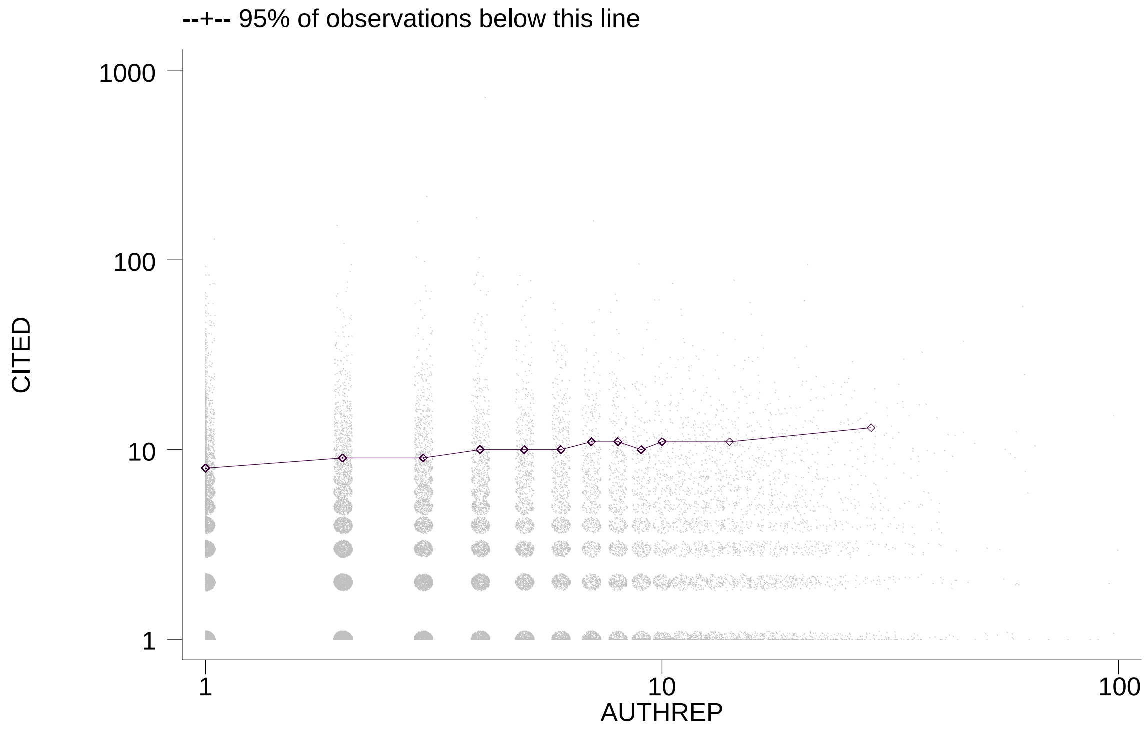

Figure 4: Relationship between CITED and AUTHREP (100,000 points sampled, points are dithered to illustrate density).

Data on the regression equation for these variables is provided in Figure 5. This equation explains 29.1% of the variance in the data, which is relatively good considering the cross-sectional nature of the data. From Table 1, we see that LJOUR only explains 22.8% of the variance (the square of the Pearson coefficient), which means the marginal improvement in adding the other two variables is relatively small (6.3%). The maximum explanatory value of author reputation is only 2.95%.

```
  Source |       SS       df       MS              Number of obs =  780049
---------+------------------------------           F(  3,780045) =       .
   Model | 142291.28       3  47430.425            Prob > F      =  .0000
Residual | 346418.81  780045  .444101063           R-squared     =  .2912
---------+------------------------------           Adj R-squared =  .2912
   Total | 488710.09  780048  .626512842           Root MSE      =  .66641

------------------------------------------------------------------------------
    LCIT |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LJOUR |   .2449059   .0008272    296.06    .000       .2432846    .2465272
   LREF  |   .1214054   .0005036    241.08    .000       .1204183    .1223924
   LAUTH |   .0731341   .0008887     82.30    .000       .0713923    .0748758
   _cons |   .0104725   .0023629      4.43    .000       .0058413    .0151038
```

Figure 5: Results from Regression Analysis (using the STATA statistical package)

Disciplinary Effects

Additional analyses were conducted to determine if journal impact measure was the most important measure in different disciplines. Table 2 lists the results for the 24 major disciplinary categories used by ISI. The list of disciplines were ordered according to the percentage of variance explained (R^2). In addition, the explanatory values of each measure are listed (the marginal explanatory value is the square of the correlation between the measure and the dependent variable). The explanatory value for an individual variable assumes that this variable is the only one explaining variance in future citation activity.

The discipline where the most variation in citing activity was explained is *Computer Science* (43% of the variance). The most important variable (journal impact) accounted for the majority of this relationship (35.8%). The discipline with the least explained variation is *Economics & Business* (20.7%). Reference impact is the most important variable in this discipline and accounts for the majority of the explained variance (14.6%).

Table 2: Variance Explained in LCITE by discipline using all three independent variables (R^2) or each variable separately. The best independent variable is shaded.

Discipline	R^2	LJOUR	LREF	LAUTH
Computer Science	0.4307	0.3586	0.2814	0.0542
Neurosciences & Behavior	0.3968	0.344	0.2958	0.0985
Geosciences	0.348	0.2708	0.2292	0.0436
Ecology/Environment	0.3358	0.2774	0.2505	0.0197
Psychology/Psychiatry	0.3333	0.259	0.2323	0.0412
Molecular Biology & Genetics	0.3097	0.2623	0.1934	0.0224
Immunology	0.3089	0.2433	0.2278	0.0518
Clinical Medicine	0.3009	0.2404	0.1919	0.0307
Mathematics	0.2992	0.2412	0.2103	0.0294
Law	0.2853	0.2095	0.1867	0.0076
Space Science	0.28	0.2123	0.2163	0.0526
Education	0.2684	0.1865	0.2125	0.0241
Multdisciplinary	0.2668	0.2053	0.1584	0.0464
Social Sciences, general	0.2621	0.1906	0.1875	0.0219
Biology & Biochemistry	0.2615	0.2052	0.1756	0.0121
Engineering	0.2609	0.1978	0.1868	0.0206
Chemistry	0.2576	0.190	0.1792	0.0329
Pharmacology	0.2567	0.1658	0.1856	0.0394
Microbiology	0.2508	0.1578	0.1998	0.0353
Agricultural Sciences	0.25	0.176	0.1813	0.0261
Materials Science	0.245	0.195	0.1819	0.0417
Plant & Animal Science	0.2316	0.1792	0.1745	0.0111
Physics	0.2294	0.161	0.1772	0.0115
Economics & Business	0.2079	0.1455	0.1457	0.0179

Overall, journal impact is most important (17 out of 24 disciplines), and is important in the disciplines which have higher variances (10 out of the top 10). Reference impact shows up as most important in some of the disciplines with lower variances (5 of the lowest 7). Author reputation has relatively low explanatory value throughout.

Discussion and Conclusions

The analysis conducted in this study supports the present policy of using the journal impact factor as a surrogate for the impact of a paper. More sophisticated methods, such as assessing the impact of the references or the past reputation of the author, are statistically significant in predicting future impacts. But these more sophisticated methods do not result in sizable improvements in predictive ability. The costs to acquire and process the additional information are probably not justified by the added benefit.

The analysis also suggests that if a paper is from a journal that doesn't have an established journal impact factor, one could use reference impact as a surrogate. Reference impact is highly inter-correlated with journal impact for the reasons mentioned earlier. Future analysis may be able to discern whether high impact journals tend to reference other high impact journals (more than low impact journals), and whether low impact journals tend to reference other low impact journals (more than high impact journals).

And perhaps most importantly, the analysis suggests that it is possible to predict a significant amount of the variance in future citation activity. The relationship holds for every discipline, and holds for science as a system. To this we add the caveat that the absolute correlation values are still low enough that one would not be justified in predicting citation counts for single papers, or even small groups of papers, in any type of research assessment exercise.

We would, however, like to suggest that results of such an analysis at the paper level could be very useful for generating semantic maps of science. Current techniques in semantic analysis (text-based approaches such as co-word, used to describe themes in the literature) often account for the relative importance of terms using normalizations such as TFIDF or log-entropy. However, these techniques also assume that all documents have an equal impact on the cognitive structure of science (Callon & Law, 1983; Callon, Law, & Rip, 1986; Noyons & van Raan, 1998).

Common sense tells us that an article published in a highly influential journal by highly influential authors and with a stellar set of references will probably have a much larger impact going forward, and thus a greater influence on cognitive thoughts in the world of research. We propose, therefore, that adjustments are needed to the weightings of documents so that these relative impacts can be taken into account. For cognitive or semantic maps of current science where articles have not had time to accrue citation counts, we suggest a document weighting scheme based on properties such as the impact factor and/or quality of references.

References

Callon, M., & Law, J. (1983). From translations to problematic networks - an introduction to co-word analysis. *Social Science Information, 22*(2), 191-235.

Callon, M., Law, J., & Rip, A. (1986). *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World.* London: The Macmillan Press.

Glänzel, W. (1997). On the possibility and reliability of predictions based on stochastic citation processes. *Scientometrics, 40*(3), 481-492.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics, 53*(2), 171-193.

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing & Management, 31*(1), 69-80.

Noyons, E. C. M., & van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68-81.

van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133-143.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics, 62*(1), 117-131.